Video Anomaly Detection based on Wake Motion Descriptors and Perspective Grids

Roberto Leyva, Victor Sanchez and Chang–Tsun Li
Department of Computer Science, University of Warwick, United Kingdom
M.R.Leyva-Fernandez@warwick.ac.uk,vsanchez@dcs.warwick.ac.uk, C-T.Li@warwick.ac.uk

***Abstract*—This paper proposes a video anomaly detection method based on wake motion descriptors. The method analyses the motion characteristics of the video data, on a video volume-by-video volume basis, by computing the wake left behind by moving objects in the scene. It then probabilistically identifies those never previously seen motion patterns in order to detect anomalies. The method also considers the perspective of the scene to compensate for the relative change in an object's size introduced by the camera's view angle. To this end, a perspective grid is proposed to define the size of video volumes for anomaly detection. Evaluation results against several state-of-the-art methods show that the proposed method attains high detection accuracies and competitive computational time.**

***Index Terms*—Video surveillance, anomaly detection, spatio-temporal video volumes, wake motion descriptor**

I. Introduction

Recently, video surveillance has become an increasingly important tool in the area of public security. To this end, a large number of cameras are currently present in many airports, subway stations and other different public places. In 2006, there were 4.2 million CCTV cameras operating in Britain and over 30 million in the USA, an amount that has increased considerable over the last years [1], [2]. Despite the large number of cameras installed, video surveillance usually involves the manual intervention of a user, which makes it in many cases infeasible. For example, in order to detect particular events at all times, one person is required per surveillance camera, and even if this is possible, human errors may hinder the detection results [3],[4]. Therefore, a number of approaches have been proposed in recent years to automatically detect unusual activities in video data [3]. These approaches can be classified into two main groups - explicit detection and anomaly detection [5]. Explicit detection approaches require to manually define the abnormal events that may occur in a particular scenario and thus the number of events that can be detected is limited [5]. For this reason, anomaly detection approaches are gaining considerable attention due to the vast number events that can be detected, and the fact that these events may be updated dynamically with little supervision. [6]-[10].

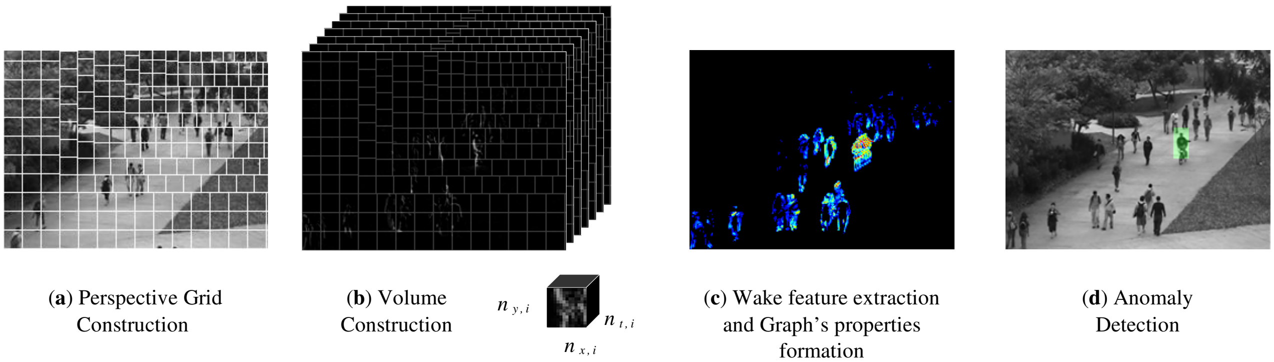
Anomaly detection approaches can be further classified into tracking and non-tracking based methods, with their respective variants of supervised and unsupervised approaches. Although both methods have been applied for anomaly detection, non-tracking based methods have provided better performance in crowded scenes mainly because it is challenging to track individuals as the crowd density increases, or to track the whole crowd as a single object [16].

Many non-tracking based methods rely on extracting and analyzing local low-level visual features from the scene [6] -[11]. In general, the aim is to detect anomalies based on the assumption that what occurs in a particular region at a particular time should be related to what has occurred in the past in the same region. To this end, a probabilistic model is defined to describe the features extracted from normal scenes. A region is then deemed to be abnormal if the corresponding features cannot be described by this probabilistic model. For example, [6] considers spatio–temporal *video volumes* within the context of Bag Of Videos words (BOV). A BOV is computed by those less similar video volumes in order to generate a probabilistic model. A volume is deemed to be similar to another as long as the Euclidean distance between them is low. New video volumes are then analyzed against the model and classified as abnormal or normal by considering the composition of the assigned video volumes using the BOV. In [9], a spatio–temporal Laplacian eigenmap method is proposed to detect anomalies within the context of crowd behavior. The method learns the spatial and temporal variations of local motions in an embedded space. A model is then constructed to characterize the normal crowd behavior. The model allows for the detection of abnormal crowd activities and the localization of regions which show abnormal behavior.

The main advantage of non-tracking based methods that rely on statistical inference processes is that they consider the spatio–temporal relations of the scene's regions in the probabilistic model; thereby the model can be updated using new observations. An important drawback, however, is the large number of frames required in the training stage to create the probabilistic model, and the associated long processing times [10]–[15]. Additionally, the majority of the methods based on video volumes, do not consider the relative position of the camera and the underlying perspective of the captured scene when defining the video volume size. Usually, a fixed voulme size is used for the whole scene independently of the object's motion and size in relation to the camera's position.

In this paper, we propose a non-tracking based method to detect video anomalies by analyzing video volumes. Our method employs a wake motion descriptor to describe the trace

This research report was financed by the National Council of Science and Technology of Mexico (CONACYT)



(**a**) Perspective Grid Construction

(**b**) Volume Construction

(**c**) Wake feature extraction and Graph's properties formation

(**d**) Anomaly Detection

Fig. 1. The proposed method. (a) A perspective grid is computed for the whole sequence taking into account the perspective of the scene. (b) Motion is detected by computing the gradient along the whole video sequence in the time direction. Video volumes are created based on the perspective grid. (c) Frames comprising each video volume are merged using the maximum operator to compute the wake motion descriptor of sequence. Based on the wake motion descriptor, a set of connected graphs are generated to describe moving objects. (d) Anomalies are detected by analyzing the graphs' properties.

left behind by a moving object in the scene. This descriptor allows decreasing the processing time by reducing the number of operations needed to extract features from video volumes. The method takes into account the perspective of the scene to construct the video volumes. Specifically, it uses varying-size volumes according to the relative position of the regions they describe in relation to the camera's position. By using a statistical inference process, the method analyzes and classifies video volumes so that those video volumes with not previously seen wakes are deemed to be abnormal.

The rest of this paper is organized as follows. In Section II we detail our method. Section III compares our method with the state-of-art methods and Section IV presents conclusions and future work.

II. Proposed method

Our method is graphically summarized in Fig. 1. First, we detect motion by computing the gradient along the whole video sequence in the time direction. We then construct video volumes taking into account the perspective of the scene, so that video volumes depicting regions close to the camera are larger than those depicting regions far from it. We then merge the frames comprising each video volume using the maximum operator to compute the wake motion descriptor of the whole scene. This descriptor is then used to generate a set of connected graphs. Finally, we infer anomalies by analyzing the graphs' properties. In the next sections, we describe each step in detail.

A. Perspective grid and video volumes

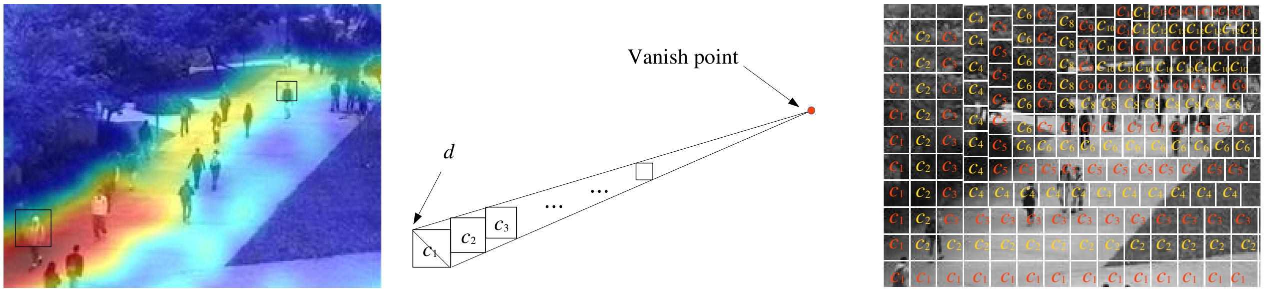
When capturing a scene, it is common that the camera's field of view be at an angled position from the main plane of movement. In this case, objects appear to move faster as they approach the camera, even if they move at a constant speed. This is illustrated in Fig. 2 (a), where the optical flow of the depicted scene increases for those regions close to the camera. Moreover, objects close to the camera appear to be larger in size than those located far from it even if they are the same size. Within the context of anomaly detection using video volumes, the relative change in an object's speed and size introduced by the camera's field of view may have a detrimental effect on the extracted features.This may be solved by scaling objects, in size and speed, as they move close to the camera in the scene; however, any scaling method, e.g. nearest-neighbor interpolation, usually requires a large number of operations for every frame in the sequence. In this work, unlike common grid methods e.g. [8], [7], we propose to create video volumes of different sizes in the X-Y plane according to the position of the regions they depict relative to the camera's position. In other words, the size of the X-Y plane of volumes increases as they depict regions that are closer to the camera's position. To this end, we create a *perspective grid* for the whole sequence, which comprises a number of cells that define the size of the video volumes in the X-Y plane. This is illustrated in Fig. 2 (c), where those cells depicting regions near the camera are larger than those located far from it. In order to construct the perspective grid, we first manually detect the vanish point p of the scene. Starting form an initial cell c_1 of size $n_{x,1} \times n_{y,1}$, we then proceed to generate smaller cells as they approach the vanish point, as illustrated in Fig. 2 (b). For each generated cell, we store its diagonal size, $\lfloor d_k \rfloor$, in the set D:

$$D = \bigcup_k \lfloor d_k \rfloor \tag{1}$$

We then use the elements in D and the location of p to sweep the scene vertically and horizontally and define the volume sizes in the X-Y plane (see Fig. 2 (c)). A video volume v_i is then obtained by selecting the subregion denoted by cell c_i of size $n_{x,i} \times n_{y,i}$ over $n_{t,i}$ fames along the sequence, as illustrated in Fig. 1 (b). In this work, $n_{t,i}$ is constant for all video volumes.

B. Wake motion descriptor

The proposed method is based on a wake motion descriptor that describes the trace left behind by a moving object in



(a) Average speed   **(b)** One – point Perspective   **(c)** Perspective Grid

Fig. 2. (a) Optical flow of a scene as it increases from blue to red for those regions close to the camera. (b) Starting form an initial cell c_1 of size $n_{x,1} \times n_{y,1}$, smaller cells are generated as they approach the vanish point of the scene. (c) Using set D, which contains the diagonal size of all cells, $D = \{\lfloor d_1 \rfloor, \lfloor d_2 \rfloor, \ldots, \lfloor d_n \rfloor\}$, the perspective grid is generated by sweeping the scene in the vertical and horizontal direction starting with cell size c_1, which is located in the point farthest from the scene's vanish point.

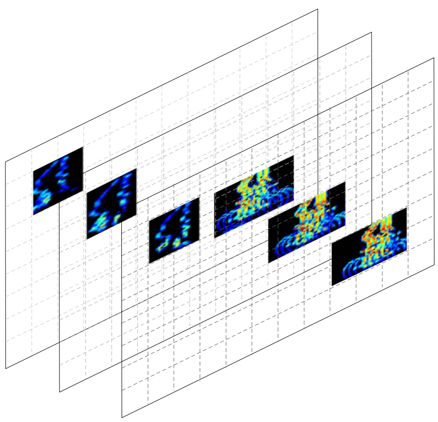


Fig. 3. Wakes: the trace left behind by two objects.

the scene and provides information about the object's motion characteristics. Fig. 3 shows the wake left behind by two objects. It is expected to have different wakes from different objects moving at different speeds, e.g. a boat crossing a lake fast or slow. Therefore, we can classify an object based on its wake. In order to calculate the scene's wake, we first compute the gradient in the time direction along the sequence:

$$M_j = I_{j+1} - I_j \tag{2}$$

where j denotes the index of frame I. For a sequence of J frames, sequence M comprises $j - 1$ gradient frames. Next, we merge n_t mini-frames inside each video volume v_i into a 2D array of size $n_x \times n_y$ to generate the corresponding wake w_i:

$$w_i = max_{\Delta t}[v_i] \tag{3}$$

where $max_{\Delta t}$ denotes the maximum operator in the time direction. We then analyze the texture in the wake w_i using the Segmentation-based Fractal Texture Analysis or SFTA descriptor. This descriptor consists in decomposing an image into a set of binary images from which the fractal dimensions of the resulting regions are computed in order to describe segmented texture patterns. SFTA provides the description of an image as a 3D vector, d, comprising the features that represent the area, mean and fractal dimension of the binarized image [17]. Using a set of normal scenes, i.e., training data, we construct a Gaussian Mixture Model (GMM) with Q Gaussian distributions for each feature d_k and determine the corresponding parameters θ_k using the Expectation Maximization algorithm. Thereby, for each feature k we have:

$$p(\theta_k|d_k) = \sum_{q=1}^{Q} N(\mu_q, \sigma_q) \tag{4}$$

where $N(\mu_q, \sigma_q)$ denotes a Gaussian distribution with mean μ_q and standard deviation σ_q. Eq. (4) is used to determine whether incoming wakes fit the model according to previously seen SFTA features of normal scenes. We consider this aspect in the graph formation criterion, as explained in the next section.

C. *Graph formation*

Small space–time regions can be represented as a set of connected graphs [9]. We then create a graph $G_i(V, E)$ comprising all those video volumes that are connected in space–time and with wakes whose SFTA features meet the following criterion:

$$\epsilon \leq -log[\prod_k p(\theta_k|d_k)] \tag{5}$$

where $k = 3$ features, ϵ is a threshold and $p(\theta_k|d_k)$ is the model in Eq. (4). In other words, if v_i and v_j are connected in space–time and their respective wake have SFTA features that satisfy (5), then an undirected edge is generated between v_i and v_j. From the generated graphs, we infer the object's size, speed and motion duration in the scene, as exemplified in Fig. 4. From a set of normal scenes, we can then infer the most common graph's properties and recognize abnormal graphs.

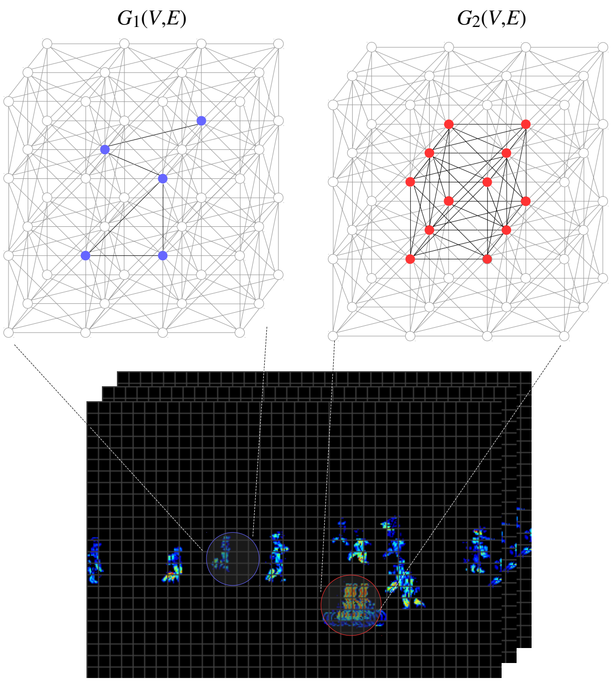


Fig. 4. Two graphs representing each the characteristics of an object's motion in the scene. The graph in blue corresponds to a pedestrian's motion while the one in red corresponds to a bicycle's motion.

D. Anomaly detection

It is possible to classify moving objects based on their corresponding graph's properties e.g. number of vertices $|V|$, number of edges $|E|$, diameter, connectivity, density, etc. In this work, we detect abnormal motion patterns by considering each graph's edge size $x_i^{|E|} = G_{i,|E|}$ and diameter $x_i^D = G_{i,D}$. The former corresponds to the longest path s in the graph for any two vertices u and v:

$$G_{i,D} = max_{u,v} s(u, v) \tag{6}$$

We classify different graphs using a normal distribution $N(\mu, \sigma)$. In the training stage, we estimate the mean, μ, and variance, σ of a set of U normal graphs. For the edge size of the normal graphs, μ and σ are given as follows:

$$\mu_{|E|} = \frac{1}{U}\sum_{i=1}^{I} x_i^{|E|} \quad \sigma_{|E|} = \frac{1}{U}\sum_{i=1}^{I}(x_i^{|E|} - \mu_{|E|})^2 \tag{7}$$

For the diameter of the normal graphs, μ and σ are given as follows:

$$\mu_D = \frac{1}{U}\sum_{i=1}^{I} x_i^D \quad \sigma_D = \frac{1}{U}\sum_{i=1}^{I}(x_i^D - \mu_D)^2 \tag{8}$$

The probability that a graph G_i fits the model is then given by:

$$p(G_i) = \prod_{k=\{|E|,D\}} \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x_i^k - \mu_k)^2}{2\sigma_k^2}} \tag{9}$$

Graph G_i is deemed to be normal if the following assumption is satisfied:

$$\gamma \geq -\log[p(G_i)] \tag{10}$$

It is important to mention that the proposed method allows incorporating future observations and updating the anomaly detection model, and most importantly, it easily allows removing graphs incorrectly deemed to be normal from the model.

III. Experimental Results

Our method is tested with the UCSD pedestrian datasets [18]. The dataset contains footages from two pedestrian sidewalks where abnormal events occur. The first scene contains 34 normal video clips and 36 abnormal video clips; the second scene contains 16 normal video clips and 14 abnormal video clips. Both scenes contain different crowd densities and non-pedestrian moving objects, namely bikes, small cars, skaters and wheelchairs, which are considered to be abnormal. We use the normal videos as training data to derive the probabilistic models in Eq. (4) and (9). For Eq. (5) and Eq. (10), a value of $\epsilon = 7.5$ and $\gamma = 2.3$ is used, respectively. For model in Eq. (4), $Q = 3$ Gaussians are used. These values were selected because they provide the best trade-off between computational complexity and detection accuracy. We evaluate our method, referred to as WMD hereafter, against a number of sate-of-art methods; specifically, Spatio–Temporal Compositions (STC) [6], Inference by Composition (IBC) [11], Laplacian Eigenmaps (LE) [9], Mixture of Dynamic Textures (MDT) [15] and Local Optical Flows (LOF) [10].

TABLE I
EQUAL ERROR RATE (EER %) FOR THE UCSD PEDESTRIAN DATASETS FOR VARIOUS METHODS AND CORRESPONDING NO. OF REQUIRED TRAINING FRAMES

Method	Dataset	Frame Level (%)	Pixel Level (%)	No. of Training Frames
WMD	Ped1	19	26	200
	Ped2	16	25	180
STC	Ped1	15	27	200
	Ped2	13	26	180
IBC	Ped1	14	26	6800
	Ped2	13	26	2880
LE	Ped1	13	22	1000
	Ped2	22	31	1000
MDT	Ped1	25	58	6800
	Ped2	24	54	2880
LOF	Ped1	38	76	6800
	Ped2	42	-	2880

There are two ways to evaluate the performance of the method, either at a pixel or frame level. The later implies considering the frame abnormal regardless of the abnormal pixel locations. In this case, it is necessary to compare the abnormal frames against the ground truth at a pixel level to locate the position of the abnormal event. We first consider a frame to be abnormal if it contains at least one anomalous pixel. For this case, the detection accuracy of the state-of-art and our proposed method is shown in the receiver operating characteristic (ROC) curve depicted in Fig. 5. Note that our method attains a very similar performance to STC and IBC,

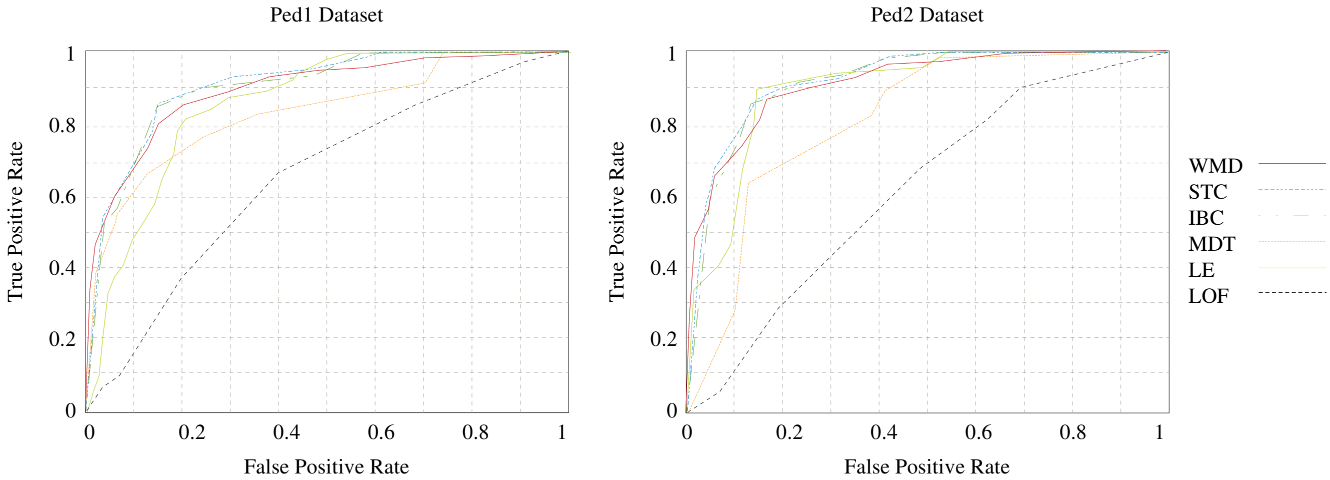


Fig. 5. Frame level ROC comparison of the proposed (WMD) and state-of-the-art methods for the UCSD pedestrian datasets.

but with less computational complexity and less number of training frames, as will be shown later.

The method is also evaluated at a pixel level. We consider a frame to be abnormal if it contains at least 40% of the abnormal pixels. Otherwise it is considered to be a false alarm. For this case, the equal error rate (EER), which denotes the percentage of misclassified frames when the false positive rate is equal to the miss rate, is reported in Table I at pixel and frame levels; while visual results are shown in Fig. 6. Results in Table I show that our method outperforms the state-of-art approaches in the UCSD pedestrian datasets in terms of detection accuracy at a pixel level. Note that our method requires the same number of frames in the training stage as STC [6], which is considered the state-of-the-art due to the small numbers of frames required to train the method. Our method also attains competitive computational times as reported in Table II. All tests were done in a PC: Intel Core Quad with 2 CPUs and 7.8GB of RAM.

TABLE II
COMPUTATIONAL TIME PER FRAME

Method	Dataset	Time (seconds)
LE	Ped1	0.0065
	Ped2	0.0065
STC	Ped1	0.19
	Ped2	0.22
WMD	Ped1	1.22
	Ped2	1.55
MDT	Ped1	21
	Ped2	29
IBC	Ped1	69
	Ped2	83

It is important to mention that although our method was able to detect all non-pedestrian movements in both datasets, it is sensitive to synchronized crowd movement in the same direction; e.g. a group of pedestrians walking in a synchronized fashion in the same direction. This is mainly due to occlusions among objects. In these cases, the wake motion descriptor may not be capable of determining whether the crowd is a combination of objects moving in the same direction at the same speed or a big moving object.

TABLE III
EQUAL ERROR RATE (EER %) OF THE PROPOSED METHOD FOR THE UCSD PEDESTRIAN DATASETS WHEN USING (WMG-G) AND NOT USING (WMD-NG) THE PERSPECTIVE GRID

Method	Dataset	Frame Level (%)	Pixel Level (%)
WMD-G	Ped1	19	26
	Ped2	16	25
WMD-NG	Ped1	26	35
	Ped2	21	27

The benefits of using the proposed perspective grid to define the size of video volumes is also evaluated. Table III reports the detection accuracy at a frame and pixel level of our proposed method when the perspective grid is used (WMD-G) and not used (WMD-NG). Here, a frame is considered to be abnormal if it contains at least 40% of the abnormal pixels in the ground truth. In both datasets, WMG-G was able to detect abnormal motion patterns in the whole frame regardless whether the moving objects are located close or far from the camera's position. The abnormal objects were detected as soon as they appear in the scene. Note that the benefits of using the perspective grid are greater for dataset Peds1 than those for UCSD dataset Peds2, since the camera in dataset Peds1 is angled from the main plane of movement, as illustrated in Fig. 6. Some aspects of our proposed method may be improved in future work. For example, we observed during the tests that a low value for threshold ϵ increases accuracy in crowded scenes. This threshold can be dynamically determined according to the number of moving objects in the scene. Another aspect is that the wake motion descriptor varies significantly and becomes a less powerful descriptor as moving objects approach the camera; thereby, the perspective of the scene can also be taken into account to define the size of video volumes in the time dimension.

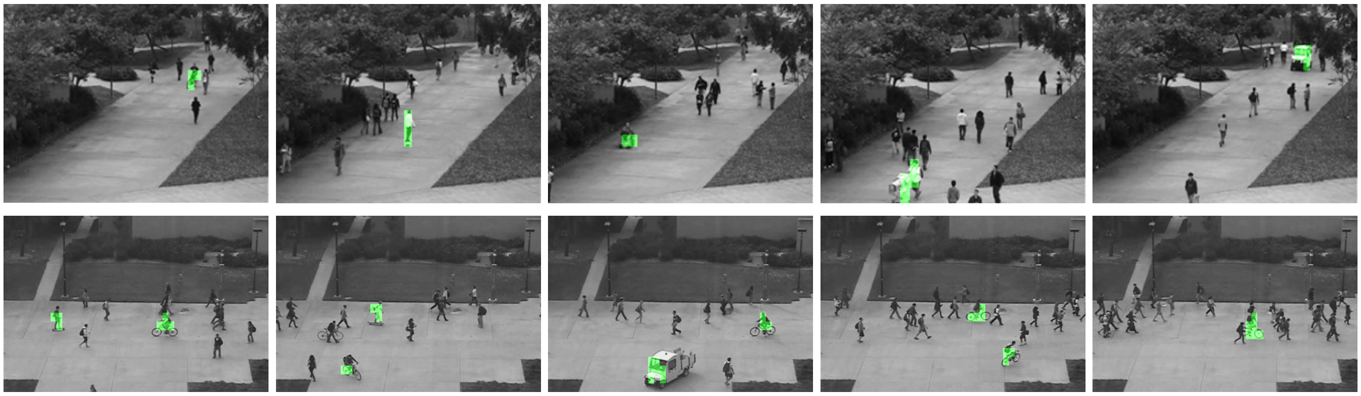


Fig. 6. Examples of abnormal frame detection for the UCSD pedestrian datasets. First row depicts scenes from the Ped1 set and second row from the Ped2 set. Abnormal regions are highlighted in green. The proposed method was able to detect cyclists, skaters, wheelchairs, and small cars.

IV. Conclusion

This paper proposed a new method for anomaly detection in video. The method is based on statistical inferences using video volumes and a wake motion descriptor. In order to compensate for the camera's position and the underlying perspective of the scene, the method uses a perspective grid to define the size of video volumes according to their proximity to the camera. Based on the wake motion descriptor, the method constructs graphs of spatio–temporal connected video volumes that describe moving objects. Those moving objects with graphs whose characteristics do not fit a probabilistic model are deemed to be abnormal. Since similarities between graphs can be obtained from an initial training stage and updated with every new frame, the proposed method is suitable for real time processing. The method was tested against several state-of-the-art methods. Results show that it provides a high detection accuracy and competitive computational complexity.

References

[1] D.M. Wood, K. Ball, D. Lyon, C. Norris, and C. Raab, "A report on the surveillance society," *Surveillance Studies Network*, 2006, UK.

[2] J. Vlahos,"Surveillance society: New hight-tech cameras area watching you," *Popular Mechanics*, pp. 64 – 69, 2008.

[3] N. Haering, P.L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 279 – 290, October 2008 .

[4] H.M. Dee, and S.A. Velastin,"How close are we to solving the problem of automated visual surveillance?," *Machine Vision and Applications*, vol. 19, no. 5-6 pp. 329 – 343, October 2008.

[5] C. Piciarelli, and G.L. Foresti, "Surveillance-oriented event detection in video streams," *IEEE intelligent systems*, vol.26, no. 3 pp. 32 – 41.

[6] M.R. Javan, and M.D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Computer Vision and Image Understanding*, vol. 117 no. 10, pp. 1436-1452, October 2013.

[7] A. Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320-329, March 2012.

[8] V. Reddy, C. Sanderson and B.C. Lovell, "Improved Anomaly Detection in Crowd Scenes via Cell-based Analysis of Foreground Speed, Size and Texture," *IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 55-62 2011.

[9] M. Thida, H. Eng, and P. Remagnino, "Laplacian Eigenmap With Temporal Constrains for Local Abnormality Detection in Crowded Scenes", *IEEE Transactions on Cybernetics*, vol. 43, no. 6., pp. 2147 – 2156, December 2013.

[10] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust Real – Time Unusual Event Detection Using Multiple Fixed – Locations Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30 no.3 pp. 555 – 560, December 2013.

[11] O. Boiman, and M. Irani, "Detecting irregularities in image and in video," *International Journal of Computer Vision*, vol.79, pp. 17 – 31, August 2007.

[12] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video." in *Proceedings IEEE Computer Society Conference on the Computer Vision and Pattern Recognition*, vol. 2, pp. II-819, 2004.

[13] R. Mehran,A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 935-942, 2009.

[14] L. Kratz, and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models." in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 1446-1453, 2009.

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pp. 1975-1981, 2010.

[16] H. Yang,L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review." *Neurocomputing*, vol.74, no. 18, pp. 3823-3831, 2011.

[17] A.F. Costa, G. Humpire-Mamani, and A.J. Machado, "An efficient algorithm for fractal analysis of textures." in *Proceedings 25th Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 39-46, 2012.

[18] UCSD Dataset. Available online: "http://www.svcl.ucsd.edu/projects/anomaly/dataset.html"